# Getting started with data

The purpose of this initial exercise is to get started with analysing the data. Download the notebook and the data file for politicians in Wikipedia, and answer the following questions. The idea of these questions is to check that everything is technically correct and you are on the right track. In the project you can try out more ideas and interpret the results.

Download the pre-fetched data here politicians.json and the template notebook here project-wikipedia.ipynb . Note that as this notebook is meant for the project, it is less structured than a typical notebook we have set up for exercises. You might need to do some exploration on how the data is formatted and write bit more code from the scratch than in a typical exercise.

Points 20 / 20 My submissions 1 / 5 Deadline Friday, 26 May 2023, 19:00 To be submitted alone

## Construct network: how many links are there

**Question 1** 20 / 20

Your first task is to construct a network of links between wikipages of the politicians. Use Graph object in NetworkX to construct an undirected network. There should be an undirected edge between nodes A and B if either there is a link from page of A to B or link from B to A, or both. Note that if you just add edges to the Graph object in Networkx, this should be automatically taken care of. Further, do not add self-edge, i.e., check that you don't add edge from a politician to themself. You are now ready to start exploring the structure of the network! How many edges does your network have?

10294

✔ Correct!

Submit

Points 20 / 20 My submissions 1 / 1 Deadline Friday, 26 May 2023, 19:00 To be submitted alone

⚠ You have used the allowed amount of submissions for this assignment.

## Network degrees by gender

**Question 1** 20 / 20

Next, compute the average degree of males and females in the network. Be sure to include nodes with degree 0, i.e., nodes that don't have any links. What do you observe?

- ◉ **Females have higher degree than males, with average degree around 7.2 as compared to male average degree around 4.1.**
- ○ Females have higher degree than males, with average degree around 5.3 as compared to male average degree around 4.5.
- ○ Males have higher degree than females, with average degree around 7.2 as compared to female average degree around 4.1.
- ○ Males have higher degree than females, with average degree around 5.3 as compared to female average degree around 4.5.

✔ Correct!

Submit

Points 20 / 20 My submissions 1 / 5 Deadline Friday, 26 May 2023, 19:00 To be submitted alone

## Summaries by gender: bag-of-words

**Question 1** 20 / 20

Your next task is to construct bag-of-words for all the summary texts of all politicians, male politicians, and female politicians. That is, you count the times each word appears in summaries of these three categories. Sort the list by frequency and inspect the most frequent ones. In all three sets of politicians (all, male, female), the three most common words should be the stopwords 1. "the", 2. "of", 3. "and". What is the **4th** most common word both in male and female summaries?

in

✔ Correct!

Submit

Points 40 / 40 My submissions 2 / 5 Deadline Friday, 26 May 2023, 19:00 To be submitted alone

## Summaries by gender: tf-idf

**Question 1** 40 / 40

Your final task is to compute the average tf-idf scores for words in each category. To do this, first compute the average tf score by simply dividing the bag-of-words number you computed in the previous question by the total number of politicians of a certain kind. For example, if there are 386 female politicians, and the word "the" appears 2293 times in their summaries, the tf value for "the for females is 2293/386. The idf score you count by first computing the document frequency df, which is the number of summary texts wherein the word appears at least once and divide that by the total number of politicians. For example, if the word "the" appears in 6149 of all 6880 summary texts then the df value is 6149/6880. The inverse document frequency score, i.e., the idf score, is then the logarithm of the inverse of the df, for example for the word "the" idf = log(6880/6149). Use the natural logarithm here. The tf-idf score is thus the multiplication of these two scores, that is, tf*idf, and its value for females and the word "the" is 2293/386 * log(6880/6149). Inspect the top tf-idf scores for males and females. For females the top score is given by "she". What is the word with **second highest** score for females?

her

✔ Correct!

Submit